基于开放域对话系统的自动化评测方法研究 *

王春柳，杨永辉，赖辉源，邓　霏

(中国工程物理研究院 计算机应用研究所，四川 绵阳 621000)

摘　要：开放域对话系统的研究在近年来取得了很大的进展，然而基于该类系统的自动化评测依然是目前亟待解决的问题。针对目前各类评测方法需要大量标注数据和评测准确率较低等问题，提出了一种利用长短期记忆网络和注意力机制判别问题-回复对是否为真实对话的评测模型。该模型基于连续的对话语料进行建模，解决了目前基于参考回复的评测模型需要大量标注数据的弊端。在 Cornell 和 Reddit 数据集上，该模型分别取得了 57.2%和 71.8%的准确率，与现有的几种评测模型相比准确率有明显提升。

关键词：对话系统；开放域；自动化评测；长短期记忆网络；注意力机制

中图分类号：TP391　　doi: 10.19734/j.issn.1001-3695.2018.10.0819

Automatic evaluation method for open domain dialogue systems

Wang Chunliu, Yang Yonghui, Lai Huiyuan, Deng Fei

(*Institute of Computer Application*, *China Academy of Engineering Physics*, *Mianyang Sichuan* 621000, *China*)

Abstract: Although great progress has been made in open domain dialogue systems in recent years, automatic evaluation methods based on these systems are still a problem to be solved. In order to solve the problem that various evaluation methods need a lot of tagged data and low accuracy, this paper proposed a model for judging whether the response pair is a real dialogue by using the long-term and short-term memory network and attention mechanism. The model was based on continuous dialogue corpus, which solves the shortcomings of the current evaluation methods based on the reference response. On the Cornell and Reddit data sets, the accuracy of the model was 57.2% and 71.8% respectively, which was obviously improved compared with the existing evaluation models.

Key words: dialogue system; open domain; automatic evaluate; long short-term memory; attention mechanism

0　引言

开放域对话系统由于其广泛的应用受到了越来越多的关注，深度学习和强化学习等方法的涌现使得该领域的研究在近年来取得了很大的进展。不同于任务型对话系统，开放域对话系统由于具有非常广泛的话题领域，导致其回复的内容具有很强的多样性和复杂性，使得目前尚未存在一个良好的方法能够实现对开放域对话系统的快速评测，这在一定程度上阻碍了开放域对话系统的长远发展。

针对开放域对话系统评测的一个非常原始的方法是采用人工评测[1]。这种方法虽然评测结果相对准确但通常开销巨大且非常耗时，评测人员的不同也使得评测结果具有很强的主观性。鉴于人工评测方法的众多缺点，研究人员陆续提出了各种自动化的评测方法。早期的自动化评测方法主要包括两类，分别为基于词重叠率的评测指标和基于词向量的评测指标[2]。基于词重叠率的评测指标以 BLUE[3]、METEOR[4]和 ROUGE[5]为代表，主要利用系统的生成回复与参考回复之间的词重叠率进行评测。基于词向量的评测指标是将系统的生成回复与参考回复表示成向量的形式，然后通过计算余弦距离来表示两者之间的相似度。虽然这些评测指标目前被工业界和学术界广泛使用，但 Liu 等人[6]通过大量的实验证明了这些评测指标与人类判断结果的相关性很低甚至没有相关性，因此使用这些方法进行的系统评测并不具备可靠性。

针对上述方法的缺点，研究学者在近年来又陆续提出了基于深度学习[7~11]和基于强化学习[13]等自动化评测方法。比较有代表性的方法是 Google 的 Kanan 等人[7]提出的对抗性评测方法，其灵感来源于图灵测试，该方法利用生成模型生成回复，再使用判别模型对生成模型生成的回复和参考回复进行区分，用于直观评价生成模型产生的回复与参考回复之间的相似程度。随后，Lowe 等人[8]提出 ADEM 模型，该模型将问题文本、生成回复和参考回复作为语料内容，以人工打分数据作为标签，使用递归神经网络模型(recursive neural network, RNN)进行自动评分模型的训练来预测生成回复的评测分数。该研究内容自发表起引起了国内外学者的广泛关注，但由于该方法需要大量的人工标注数据，所以并不具备很好的灵活性和扩展性。

为避免使用大量的参考回复和人工打分数据，Lowe 等人[9,10]提出使用无标注的对话数据作为训练语料,将长短期记忆网络模型(long short-term memory，LSTM)作为评测模型预测生成回复是真实回复的概率。该方法主要针对目前开放域对话系统领域训练数据匮乏的难题，减少了评测工作所需的巨大成本，其灵感来自于语篇连贯领域的研究。本文受这类评测方法的启发，提出 AB-LSTM-*bi*-MLP 评测模型，该模型为判别模型，主要基于 Bi-LSTM 网络模型和注意力机制(attention mechanism)，同时引入 Severyn 等人[14]提出的二次特征(quadratic feature)方法，利用多层感知机(multi-layer perceptron，MLP)预测问题—回复对为真实对话的概率值。

收稿日期：2018-10-02；**修回日期**：2018-12-10　**基金项目**：国防基础科研计划重点项目（JCKY2016212B004）

作者简介：王春柳（1993-），女，吉林辽源人，硕士研究生，主要研究方向为语义计算、对话系统评测（spring_willow@163.com）；杨永辉（1973-），男，江西丰城人，研究员，博导，主要研究方向为信息技术；赖辉源（1992-），男，海南万宁人，硕士研究生，主要研究方向为情感分析、自动问答；邓霏（1988-），男，四川绵阳人，工程师，主要研究方向为软件仿真测试.

1 相关知识

1.1 LSTM 模型

标准的递归神经网络无法较好地处理长距离依赖信息，随着文本序列间隔的增大，很容易出现梯度消失的问题。为解决这一问题，Hochreiter 等人[15]在 1997 年提出了一种特殊的递归神经网络 LSTM，它在 RNN 的基础上引入了门控机制，通过门控机制来决定对某个信息的记忆或遗忘，使得模型具有了更好的记忆功能，更加适用于处理和预测时间序列中间隔和延迟比较长的重要事件。LSTM 的具体计算公式[16]如式(1)~(6)所示。

$$i_t = \sigma(x_t U^i + h_{t-1} W^i + b_i) \tag{1}$$

$$f_t = \sigma(x_t U^f + h_{t-1} W^f + b_f) \tag{2}$$

$$o_t = \sigma(x_t U^o + h_{t-1} W^o + b_o) \tag{3}$$

$$q_t = \tanh(x_t U^q + h_{t-1} W^q + b_q) \tag{4}$$

$$p_t = f_t * p_{t-1} + i_t * q_t \tag{5}$$

$$h_t = o_t * \tanh(p_t) \tag{6}$$

其中：i_t、f_t 和 o_t 分别代表 LSTM 有三个门装置，即输入门、遗忘门和输出门，这三个门的输出都是通过将当前时刻 t 的输入 x_t 和前一时刻 $t-1$ 的输出 h_{t-1} 组合后再经过一个 sigmoid 函数得到；σ 即表示 sigmoid 激活函数；W 和 U 表示权重矩阵；b 表示 LSTM 在训练过程中学习到的偏置值；p_t 代表 LSTM 的记忆单元；h_t 则表示 t 时刻根据输出门 o_t 和 p_t 计算得来的隐藏层。

Bi-LSTM 网络模型是受双向 RNN 的启发，采用了两个方向的 LSTM 网络，其处理过程是在正向传播的基础上再进行一次反向传播，最后将这两个网络连接同一个输出层。该模型既能够保存上文信息，也能够考虑下文信息，解决了单向 LSTM 只记录上文信息的弊端。Bi-LSTM 的计算公式如式(7)~(9)所示。

$$\overrightarrow{h_t} = \overrightarrow{\text{LSTM}}(h_{t-1}, x_t, p_{t-1}), t \in [1, T] \tag{7}$$

$$\overleftarrow{h_t} = \overleftarrow{LSTM}(h_{t+1}, x_t, p_{t+1}), t \in [1, T] \tag{8}$$

$$H_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{9}$$

其中：H_t 表示 Bi-LSTM 在 t 时刻的隐藏状态。

1.2 注意力机制

在认知科学中，人类会选择性地关注信息的一部分内容，而忽略其他可见的信息，这种机制即为注意力机制。注意力机制起源于人类视觉领域的研究，Google 团队[17]在 2014 年首次提出在 RNN 模型上使用注意力机制来对图像进行分类。随后，Bahdanau 等人[18]提出将其应用到机器翻译任务中，其显著的效果使得注意力机制在自然语言处理(natural language processing, NLP)领域引起了广泛的关注，随后越来越多的学者将注意力机制引入到神经网络模型中来处理各种 NLP 任务。

在自然语言处理领域，注意力机制的主要作用体现在对文本语义的抽取。大量的研究成果说明了引入注意力机制的模型相比于单独的模型能够更好地捕捉到影响两个句子之间的连贯性或关联度的词语。注意力机制在近年来已产生了多种变体，如加性注意力、点积注意力、自注意力[19]等。本文所使用的自注意力机制的计算公式如式（10）~（12）所示.

$$e_t = u * \tanh(WH_t + b) \tag{10}$$

$$a_t = \frac{\exp(e_t)}{\sum_{k=1}^{n} \exp(e_k)} \tag{11}$$

$$v = \Sigma_t a_t H_t \tag{12}$$

其中：e_t 表示计算注意力概率权重值的中间过渡值；u、W 和 b 表示可学习的网络参数；n 表示输入数据的元素总数；a_t 表示分配到第 t 个元素上的注意力；v 表示经过注意力机制处理后的输出特征向量。

1.3 多层感知机

多层感知机是一种前向结构的人工神经网络，由输入层、隐藏层和输出层组成，其中隐藏层可以设置为多层，每一层全连接到下一层，因此也常被称为全连接神经网络(fully connected network，FCN)，其作用是对一组输入向量进行非线性拟合得到一组输出向量。

多层感知机的第一层表示输入层，负责接收信息，如输入一个 n 维向量，就有 n 个神经元。隐藏层神经元负责对输入信息进行加工处理，假设输入层使用向量 X 来表示，则隐藏层输出的计算形式为

$$Y_2 = f(W_1 X + b_1) \tag{13}$$

其中：W_1 表示权重矩阵；b_1 表示偏置项；$f(\cdot)$ 表示激活函数。

对于第 l 层，L_l 表示该层的所有神经元，其输出为 Y_l，其中第 i 个节点的输出为 $y_l^{(i)}$。连接第 l 层和第 $l+1$ 层的权重矩阵为 W_l，第 l 层第 i 个节点到第 $l+1$ 层第 j 个节点的权重为 $w_l^{(ij)}$。b_l 表示第 $l+1$ 层的偏置项，不同节点的偏置量可能不一样，第 $l+1$ 层第 i 个节点的偏置项为 $b_l^{(i)}$。最简单的 MLP 只包含一个隐藏层，即三层结构，如图 1 所示，其输入层所表示的输入信息为 $X = [x_1, x_2, x_3]$，输出层共包含两个神经元，表示的输出信息为 $Y = [y_3^{(1)}, y_3^{(2)}]$。

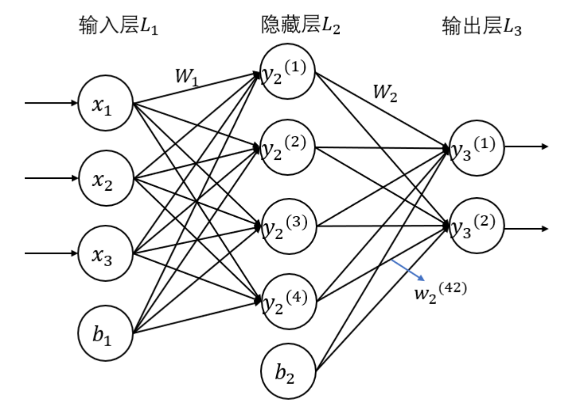


图 1 多层感知机模型图

Fig. 1 Multilayer perceptron model diagram

2 评测模型设计

本文提出的评测模型 AB-LSTM-*bi*-MLP 的整体结构如图 2 所示。该模型共包括两部分：第一部分由 Bi-LSTM 模型和注意力机制组成，用于获取文本的句子向量表示，本文将其称为句子模型；第二部分引入了“二次特征”方法，利用多层感知机预测问题—回复对为真实对话的概率值，本文将这部分称为文本对匹配模型。

句子模型可分为三层，分别为词嵌入层(word embedding)、Bi-LSTM 层、注意力机制处理层。

2.1 句子模型

词嵌入层也可以称为模型的输入层，其主要作用是将文本处理成模型可以接受的形式。其具体内容是将句子 s 看做由词语组成的序列 $[x_1, x_2, \ldots, x_n]$，每个词语都可以从词汇表 V 中提取。词语使用分布式向量 $\boldsymbol{x}$ 来表示，该向量通过在词嵌入矩阵 W 中查找来获取，矩阵 W 是由词汇表 V 中所有词语的向量表示级联而形成。为便于在 W 中快速查找到词语的向量表示，每

个词语都被映射到一个整数索引上，索引与词汇表 V 中位置对应。由此，对于每个输入句 s，可以构建一个句子矩阵 $S \in R^{n*d}$，矩阵中的第 i 行表示在句子相应位置 i 处的词语的词嵌入 x_i。

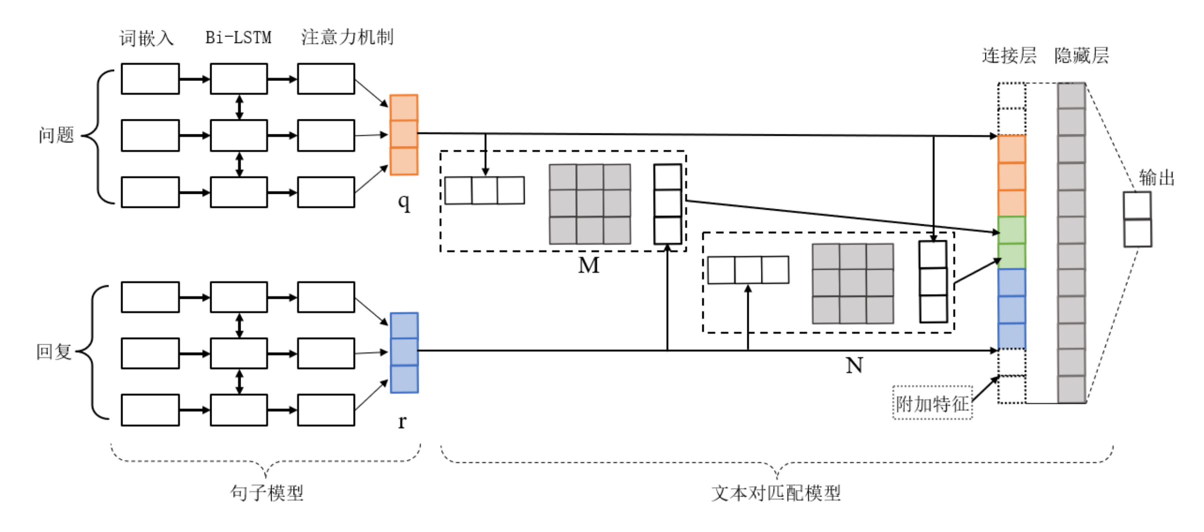


图 2　评测模型整体结构

Fig. 2　Overall structure diagram of evaluation model

Bi-LSTM 层构建了双向的 LSTM 网络，然后将词嵌入层获取到的句子矩阵输入到网络中。如图 3 中 Bi-LSTM 模型(上半部分)所示，对于每个词向量 x_i，通过式(1)~(6)可以得到在 LSTM 单元中 i 时刻的输入门、输出门、遗忘门三个门的值和整个单元的隐藏状态 h_i。通过式(7)和(8)分别获取 i 时刻的正向隐藏状态 $\overrightarrow{h_i}$ 和反向隐藏状态 $\overleftarrow{h_i}$，然后利用式(9)得到 i 时刻完整的隐藏状态 H_i，将 H_i 作为 Bi-LSTM 层输出的特征向量。

注意力机制处理层是指在 Bi-LSTM 模型的基础上再引入注意力机制，引入后的模型整体结构如图 3 所示。经过 Bi-LSTM 层处理后的输入集合可以表示为 $H=[H_1,H_2,\ldots,H_i,\ldots,H_n]$，$H_i$ 表示输入集合中第 i 个词语的特征向量，其注意力概率权重值 a_i 通过式(10)和(11)获得，将特征向量 H_i 与权重值 a_i 的乘积作为第 i 个词语的向量值 v_i，通过式(12)得到整个输入文本的输出特征向量。该层的主要作用是通过计算注意力概率分布来突出句子中的关键词语。

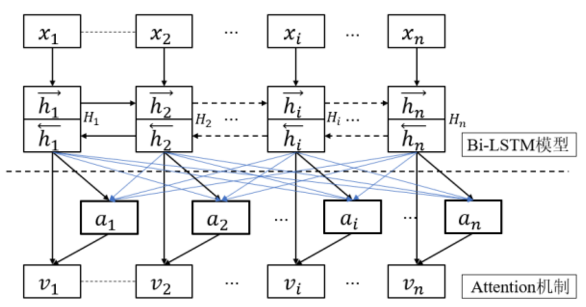


图 3　引入注意力机制的 Bi-LSTM 的模型图

Fig. 3　Model diagram of Bi-LSTM with attention mechanism

2.2　文本对匹配模型

文本对匹配模型是图 2 所示的右侧部分。该模型共包括三部分内容：a)将问题—回复对之间的匹配度作为文本对的二次特征；b)将二次特征与文本对特征向量进行连接得到连接层；c)将连接层作为多层感知机的输入预测问题—回复对是真实对话的概率值。

使用上述的句子模型对问题文本和回复文本进行处理后，得到它们的句子向量表示 q 和 r，使用式(14)[9,11]来计算问题-回复对之间的匹配度。

$$Mat(q,r)=q^T Mr \tag{14}$$

其中：矩阵 $M \in R^{n*n}$ 表示一个参数矩阵，在训练过程中不断进行优化，通常 M 被解释为将回复内容映射到问题空间的线性投影。该方法目前被广泛应用到文本相似度计算中 [14]。

由于在连贯的句子对中，机器不仅能够根据上文得出下文，也能够根据下文得到上文。因此，为了强调问题-回复对之间的双向依赖关系，本文使用式(15)表示从下文到上文的关联程度。

$$Mat(r,q)=r^T Nq \tag{15}$$

通过参数矩阵 M 和 N 分别得到问题—回复对的正向匹配度 $q^T Mr$ 和反向匹配度 $r^T Nq$，然后将其作为新的文本特征与问题 q 和回复 r 进行连接，得到整个模型的连接层，与此同时也可以选择性地在连接层中加入附加特征[14]。连接层最后的表示公式如式（16）所示。

$$X_{join}=[q,x_{mat},x'_{mat},r,x_{feat}] \tag{16}$$

其中：x_{mat} 表示 $\mathrm{Mat}(q,r)$；x'_{mat} 表示 $\mathrm{Mat}(r,q)$；x_{feat} 表示可以选择添加到连接层的附加特征。

将得到的连接层输入到多层感知机中。本文所使用的多层感知机模型的隐藏层设置为一层，激活函数使用 ReLU，输出层的激活函数使用 Softmax。模型的损失函数使用交叉熵损失函数，其定义如式（17）所示。

$$\mathcal{L}=-\Sigma_i\,\mathrm{label}_i\log(\mathrm{score}_i) \tag{17}$$

其中：$label_i$ 表示期望输出；$score_i$ 表示实际输出。

模型的反向传播使用自适应矩估计 (adaptive moment estimation，Adam)[20]优化算法。为防止模型出现过拟合现象，本文采用两种技术：Dropout 和 L2 正则化。

3　实验与分析

3.1　实验数据

为了验证本文提出的模型的有效性，本文选用两种常用的数据集对模型进行验证，分别为 Cornell 数据集(http://www.cs.cornell.edu/~cristian/Cornell_Movie-Dialogs_Corpus.html) 和 Reddit 数据集 (https://www.kaggle.com/reddit/reddit-comments-may-2015/home)。Cornell 数据集是由康奈尔大学从原始的电影剧本中所提取的对话集合，Reddit 数据集则主要由 Reddit 论坛上的评论数据构成。本文直接使用 Github 上开源的 Cornell 和 Reddit 两种数据集 (https://github.com/bshao001/ChatLearner/tree/master/Data)，分

别包含约 4 万个问题-回复对和 11 万个问题-回复对，均为连续的对话数据，其数据集样例如表 1 所示。

表 1　数据集样例

Table 1　Dataset sample

问题文本	回复文本
Daddy, people expect me to be there!	If Kat 's not going, you 're not going.
My PSU does not have that switch.	Sorry, the switch is on another similar PSU from EVGA.
Now you know who and what Freddy really is.	I though Freddy was just an old town story.

为了实现对模型的有效训练，本文将 Cornell 数据集和 Reddit 数据集中的问题-回复对标注标签值 1 作为正样本，然后在数据集中随机替换回复文本作为问题文本的回复，并将其标注标签值 0 构成模型所需的负样本。通过这种负采样方法获取与正样本数目相同的负样本，最后这两种数据集分别包含了 8 万个问题-回复对和 22 万个问题—回复对。

从 Cornell、Reddit 两种数据集中分别抽取 5%的样本对作为各自的测试集，其中 50%为真实对话，50%为虚假对话。将剩余的数据集分别按照 90%和 10%的比例分为训练集和验证集。

3.2　模型参数设置

词向量使用 word2vec 工具预训练新闻语料得到的 300 维词向量，解压后的文件大小为 3.39 GB。对未出现的词进行随机初始化，训练过程中词向量动态更新。

Bi-LSTM 网络中的正向和反向隐藏层的节点数均设置为 300 个，MLP 的隐层设置为一层，隐层节点数设置为 1 024，dropout 设置为 0.5，L2 正则化参数为 0.001，Adam 的学习率设置为 0.000 1，对于误差的更新采用批处理形式，batch 值设置为 128。

3.3　实验结果与分析

使用上述的两种数据集对模型进行训练，然后将训练好的模型分别对测试集进行预测。为验证模型的有效性，将本文模型 AB-LSTM-*bi*-MLP 与其他六种模型进行对比实验。这六种模型的结构描述如下，其中，模型名称中的“*bi*”表示使用双向的问题—回复匹配度，“*uni*”表示使用单向的问题—回复匹配度。

①AB-LSTM-*uni*-MLP：相对于本文模型仅使用单向问题—回复对匹配度作为二次特征。

②Bi-LSTM-*bi*-MLP：相对于本文模型未引入注意力机制。

③GRU-*uni*-MLP：由 Tao 等人[11]提出的无参考回复评价模型(unreferenced metric blended evaluation routine)。使用 GRU(gated recurrent unit)模型对文本进行编码，引入了单向的问题-回复对匹配度作为二次特征。

④AB-LSTM-*bi*-SLP：相对于本文模型仅使用单层感知机(Single Layer Perceptron，SLP)进行概率预测。

⑤AB-LSTM-MLP：由 Bruni 等人[10]提出的评价模型。使用 Bi-LSTM 对文本进行编码，利用 MLP 模型预测问题-回复对是真实对话的概率，相对于本文模型该模型未引入二次特征方法。

⑥LSTM-*uni*：由 Lowe 等人[9]提出的评价模型，使用 LSTM 模型对文本进行编码，直接使用问题-回复对之间的匹配度作为文本对为真实对话的概率值。

模型的评价指标选择准确率(accuracy)、精确率(precision)、召回率(recall)和 F1 值(F1 Measure)。实验结果如表 2 所示。对实验结果（表 2）分析如下：

a) 通过模型简化测试(ablation test)证明了本文模型的每一部分都对整体的评测效果起到一定的作用。

①本文模型的评测准确率优于 Bi-LSTM-*bi*-MLP 模型的评测准确度，证明了引入注意力机制后的模型能够更好地捕捉到问题-回复对之间的关联信息。

②本文模型与 AB-LSTM-*uni*-MLP 模型的评测准确率在两种数据集上均高于 AB-LSTM-MLP 模型，证明了本文模型所引入的二次级联方法的有效性。

③本文模型的评测准确率优于 AB-LSTM-*uni*-MLP 模型的评测准确率，说明了本文模型所提出的双向问题—回复对匹配度计算方法的可行性。

④本文模型的评测准确率优于 AB-LSTM-*bi*-SLP 模型的评测准确率，验证了本文所使用的 MLP 模型预测概率值方法的优越性。

b)基于 Cornell 语料的评测准确率普遍低于基于 Reddit 语料的评测结果，其原因主要与语料自身的特点有关。该语料相对 Reddit 语料数据较少；语料均来自于电影对话，噪声非常大；对话内容生僻，不如 Reddit 语料的对话内容自然。这些因素使得该语料能够覆盖的语义非常有限，导致模型很难对数据进行有效地拟合。

c)本文的评测方法在评测任务中取得了优于其他几种评测方法的判别准确率，但依然不能取得非常理想的评测结果，其原因包括以下两点：

①实验数据负采样部分所使用的随机替换方法虽然避免了数据标注所需的大量工作，但能够覆盖的语义空间非常有限，导致在较大的开放域中效果会受到影响。

②一个潜在限制是负样本中的回复文本是从数据集中的其他位置抽样而来，这些回复文本也可能与问题文本之间是连贯的，因此在一定程度上会影响模型的训练效果。

表 2　实验结果

Table 2　Experimental result

模型	Cornell				Reddit			
	准确率	精确率	召回率	F1 值	准确率	精确率	召回率	F1 值
LSTM-*uni*	0.2857	0.4907	0.2592	0.3365	0.3764	0.3513	0.3679	0.3587
AB-LSTM-MLP	0.5150	0.4974	0.5162	0.5050	0.5109	0.5391	0.5095	0.5220
AB-LSTM-*bi*-SLP	0.5233	0.5307	0.5232	0.5248	0.6861	0.7461	0.6645	0.7013
GRU-*uni*-MLP	0.5425	0.5441	0.5434	0.5420	0.6738	0.7134	0.6615	0.6837
Bi-LSTM-*bi*-MLP	0.5445	0.5086	0.5490	0.5276	0.6554	0.6689	0.6509	0.6578
AB-LSTM-*uni*-MLP	0.5543	0.5524	0.5805	0.5649	0.7089	0.7085	0.7080	0.7067
AB-LSTM-*bi*-MLP	**0.5722**	0.5688	0.5740	0.5703	**0.7182**	0.7178	0.7173	0.7158

4　结束语

本文针对开放域对话系统的自动化评测提出了一种二分类的评测模型，该模型主要结合长短期记忆网络和注意力机制，在不需要使用参考回复和人工打分数据的条件下，相比于前人提出的几种基于参考回复的评测方法更加容易实现，并且可以移植到不同的对话领域和语言上，具备很好的灵活性和扩展性。

本文的主要工作是基于对话系统中单轮的对话数据进行评测，随着近年来多轮对话系统的兴起，多轮对话系统评测也将是未来重要的发展方向。今后的研究工作将集中于多轮对话中语意连贯性的建模，通过从对话语义和对话主题等不同角度对多轮对话系统评测展开研究，同时将本文模型应用到多轮对话系统评测中，以便于进行实验对比和模型改进。

参考文献：

[1]　Curry A C, Hastie H, Rieser V. A review of evaluation techniques for

social dialogue systems [C]// Proc of the 1th ACM SIGCHI International Workshop on Investigating Social Interactions with Artificial Agents, 2017.

[2] 张伟男，张杨子，刘挺. 对话系统评价方法综述 [J]. 中国科学：信息科学, 2017, 47 (8): 953-966. (Zhang Weinan, Zhang Yangzi, Liu Ting. Survey of evaluation methods for dialogue systems [J]. Scientia Sinica Informationis, 2017, 47 (8): 953-966.)

[3] Papineni K, Roukos S, Ward T, *et al.* BLEU: a method for automatic evaluation of machine translation [C]// Proc of the 40th Annual Meeting on Association for Computational Linguistics. 2002: 311–318.

[4] Banerjee S, Lavie A. METEOR: an automatic metric for mt evaluation with improved correlation with human judgments [C]// Proc of ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization. 2005.

[5] Lin C Y. Rouge: a package for automatic evaluation of summaries [C]// Proc of Workshop on Text Summarization Branches Out. 2004: 25–26.

[6] Liu C W, Lowe R, Serban I V, *et al.* How not to evaluate your dialogue system: an empirical study of unsupervised evaluation metrics for dialogue response generation [C]// Proc of the 21th Conference on Empirical Methods in Natural Language Processing, 2016: 2122-2132.

[7] Kannan A, Vinyals O. Adversarial evaluation of dialogue models [R/OL]. https://arxiv. org/pdf/1701. 08198. pdf.

[8] Lowe R, Noseworthy M, Serban I V, *et al.* Towards an automatic turing test: learning to evaluate dialogue responses [C]// Proc of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 1116-1126.

[9] Lowe R, Serban I V, Noseworthy M, *et al.* On the evaluation of dialogue systems with next utterance classification [C]// Proc of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2016: 264-269.

[10] Bruni E, Fernandez R. Adversarial evaluation for open-domain dialogue generation [C]// Proc of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2017: 284-288.

[11] Tao Chongyang, Mou Lili, Zhao Dongyan, *et al.* RUBER: an unsupervised method for automatic evaluation of open-domain dialog systems [J/OL]. https://arxiv. org/pdf/1701. 03079. pdf.

[12] Sugiyama H, Meguro T, Higashinaka R. Automatic evaluation of chat-oriented dialogue systems using large-scale multi-references [C]// Advanced Social Interaction with Agents. 2019: 15-25.

[13] Rieser V, Lemon O. Reinforcement learning for adaptive dialogue systems [M]. Berlin : Springer, 2011.

[14] Severyn A, Moschitti A. Learning to rank short text pairs with convolutional deep neural networks [C]// Proc of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2015: 373–382.

[15] Hochreiter S, Schmidhuber J. Long short-term memory [J]. Neural Computation, 1997, 9 (8): 1735-1780.

[16] Yin Wenpeng, Kann K, Yu Mo, *et al.* Comparative study of CNN and RNN for natural language processing [J/OL]. https://arxiv. org/pdf/1702. 01923.

[17] Mnih V, Heess N, Graves A, *et al.* Recurrent models of visual attention [C]// Proc of the 27th International Conference on Neural Information Processing Systems. 2014: 2204-2212.

[18] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. Computer Science, 2014.

[19] Lin Zhouhan, Feng Minwei, Santos C N D, *et al.* A structured self-attentive sentence embedding [C]// Proc of the 5th International Conference on Learning Representations. 2017.

[20] Diederik P K, Jimmy L B. Adam: A method for stochastic optimization [C]// Proc of the 3th International Conference on Learning Representations. 2015.